DMPedia Lecture Notes in Computer Science & Engineering
Volume 1, pp. 398-408 2026

International Conference on Multidisciplinary Perspectives in Advanced Computing and Technology (IMPACT 2026)
G. B. Pant University of Agriculture and Technology, Uttarakhand, India. Jan. 10-11, 2026

Recognition of Image Caption Generation Using Deep Neural Framework

Kola Navya[1], Vanjari Masteyar Sampath Kumar[2], Dumpati Santhosh Kumar[3], Aree Rana Prathap[4]

[1,2,3,4]Department of Computer Science and Engineering, Vijay Rural Engineering College, Rochis Valley, Manikbhandar, Telangana, India

navya.mettu92@gmail.com, encryptionboys11@gmail.com

ABSTRACT

Computer vision and natural language processing techniques are required to generate textual descriptions of a given image, a task known as image captioning. Over the past ten years, research on Deep Learning and Neural Networks has upsurge as a result of the improved results. Nowadays, image processing is the main way to gather information from images, process them for a purpose, and performing operations on them. It also helps in finding a lot of information from a single image. This paper presents Recognition of Image Caption Generation Using Deep Neural Framework. By Flickr 8k dataset is used to experimentally evaluate the proposed methods. The goal of this paper is to use deep learning to detect, produce, and recognise meaningful captions for a given image. The Principal Component Analysis (PCA) method is used to extract the image's features. In this paper, RNN is used to detect, identify images, and generate captions from them. This model is evaluated based on standard evaluation metrics such as performance Time, Loss, prediction Accuracy, and BLEU (Bilingual Evaluation Understudy). Experimental results show that the described model achieves impressive performance compared to strong baseline methods.

Keywords: *Image processing, Deep Learning, Image captioning, Recurrent Neural Network (RNN), Flickr 8k, Principal Component Analysis (PCA).*

I. Introduction

The need for analytics of image data to design effective information processing systems is growing, as images are widely used to convey enormous amounts of information through social media and the internet [1]. Systems can therefore automatically analyse the content of an image and express it in meaningful sentences using natural language [2].

In recent years, image captioning has emerged significant research area. because of its potential uses in a variety of domains [3]. Using natural language to automatically generate Image captions is a 5process of describing an image's visual content. Understanding an image means identifying objects and knowing how they relate to each other [4]. The relationship determines the sequence in which the words should be arranged in a sentence and corresponds to other linguistic constituents (such as verbs). The detected objects match the nouns in the caption that need to be generated. The ability to generate natural descriptions of images is highly beneficial for practical applications such as text-based image retrieval, human-robot interaction, and assisting the blind and visually impaired [5].

Generating accurate captions for images has remained a significant challenge in artificial intelligence, despite its many uses, ranging from robotic vision to helping the visually impaired [6]. Providing precise subtitles for videos in scenarios like security systems is long-term applications also involve. “Image caption generator” [7]: As the name suggests, our goal is to build the best possible system that can generate image captions that are both grammatically and semantically accurate. As researchers seek an effective approach to improve predictions, several methods have achieved good results [8].

Handcrafted features and conventional machine learning algorithms were heavily relied on in early approaches to image captioning [9]. To generate captions, these techniques often involve extracting low-level visual features from images, such as texture descriptors or colour histograms, and combining them with linguistic models [10]. However, these early systems had struggled to capture the complex semantics and contextual understanding needed to generate informative and coherent descriptions [11].

One of the most popular developments in artificial intelligence and ML in recent years is deep learning, a technique for ML that draws inspiration from the human brain. Since deep learning has revolutionized field by providing powerful frameworks for both natural language processing and image understanding it plays a role in image captioning [12]. It uses algorithms like convolutional neural network, RNN long short-term memory, etc., Visually impaired people's lives could be improved in areas where many developments have already been made for them. To identify the objects and information in an image, for instance, voice-based image caption generator is utilized [13].

This paper presents Recognition of Image Caption Generation Using Deep Neural Framework. RNN are used in this paper for detection, recognition and generating captions from images. RNNs can capture temporal dependencies between words in a sentence, making them well suited for modeling sequential data. When it comes to image captioning, RNNs effectively bridge gap between visual perception and linguistic comprehension by decoding visual features that PCA generate corresponding textual descriptions. Using standard evaluation metrics such as Time, Loss, Prediction Accuracy, and BLEU, the model's performance is evaluated.

The remaining paper is organised as follows: Section II presents the literature survey, and Section III presents the Recognition of Image Caption Generation model. Results and discussions are elaborated in Section IV. Finally, the paper concludes with Section V.

II. Literature Survey

In [14], a novel approach of using Contrastive Language–Image Pre-training (CLIP) encodings as image features for an LSTM-based textual decoder model trained on the COCO dataset was introduced. To evaluate the model, we have generated captions for random unseen images. The model performed well with the unseen images, generating meaningful captions related to the image.

In [15], a new approach to image captioning, cross-modal prediction and relation consistency (CPRC), is proposed, aiming to constrain the generated sentence in the semantic space using raw image input. Under complex nonparallel scenarios, results demonstrate that our technique outperforms state-of-the-art comparison approach on Microsoft: Common Objects in Context (MSCOCO) "Karpathy" offline test split. On the CIDEr-D score, for instance, CPRC achieves a gain of at least 6%.

In [16], our proposed image captioning method uses real, synthetic, and real data for training and testing. To produce synthetic images, we use a text-to-image generator that is based on generative adversarial networks (GANs).

We use an attention-based image captioning method for both synthetic and real images in order to generate captions. Our experimental results show that our suggested work has two advantages: i) it shows the effectiveness of image captioning for synthetic images, ii) it further improves the quality of generated captions for real images, which makes sense given that we use more images for training.

In [17], a novel framework for remote sensing image captioning (RSIC) is proposed that uses explainable words and sentences. The word extractor and sentence generator are two components of the suggested word–sentence framework. First, it extracts valuable words from a remote sensing image, then organises them into a well-formed sentence. We then conduct comparative experiments on these 3 data sets to objectively evaluate the proposed word–sentence framework, and the results are comparable to those of encoder–decoder-based approaches.

In [18], a baseline image captioning model based on Hierarchical Attention Fusion (HAF) and reinforcement learning (RL) is introduced, combining hierarchical attention with multi-level ResNet feature maps. A scoring network (SN) is implemented to assign a score to a batch of captions in order to generate sentences based on the corresponding ground truth. A sentence-level reward, this award can benefit from more, unparalleled ground truth. Suggested approaches have achieved competitive performance when compared to similar picture caption methods, according to experimental results on COCO dataset.

In [19], a new image captioning technique is proposed that consists of 4 modules: an enhanced topic predictor (ETP), a caption generation module, a retrieval-based topic reweighting module (RTR), and a subsequent topic predictor (STP). The generation and prediction modules are trained end-to-end to handle effective topic usage by predicting appropriate topics at each time step. Comprehensive experiments on the MS-COCO and Flickr30K benchmark datasets demonstrate that our approach outperforms several recent image captioning approaches across evaluation criteria and improves the performance of the topic prediction task.

In [20], three fundamental parts of the suggested vision-enhanced and consensus-aware transformer (VCT) are a consensus-aware knowledge representation generator, a consensus-aware decoder, and a vision-enhanced encoder. VCT employs both visual information and consensus knowledge for image captioning. To learn consensus-aware representations, statistical

co-occurrence among semantic concepts is computed, and a word-correlation graph is constructed. Results from experiments on two well-known benchmark datasets demonstrate that our suggested approach achieves at state-of-the-art. Each component's effectiveness is also validated by extensive ablation studies.

In [21], propose Noise Augmented Double-Stream Graph Convolutional Networks (NADGCN) that improve the generalisation of a language model by novelly using additional background context. Based on the recipe, a rescaled grid graph can encode relationships between grid regions in the full image rather than on significant regions, NADGCN technically uses a grid-stream GCN as a supplement to the region stream. Grid-stream GCN and noise-agent designs have been validated through extensive experiments on MSCOCO, and our generator clearly outperforms the comparison baselines.

In [22], a domain-specific image caption generator is presented that uses object and attribute information to generate captions via an attention mechanism. It then uses a semantic ontology to reconstruct the generated caption and generate a natural-language description for the specified domain. To demonstrate the efficiency of the suggested model, we use the MSCOCO dataset to qualitatively and quantitatively evaluate the image caption generator.

In [23], the goal is to evaluate a comprehensive set of 6 of 7 methods, including 6 existing approaches from prior research and 1 newly proposed approach. These approaches are trained and evaluated on the Flickr8K dataset with bilingual evaluation (BLEU). With a BLEU-1 score of 0.532143 and a BLEU-4 score of 0.126316, the suggested ResNet50 – BERT – Bahdanau Attention model performs better than other models in our tests.

In [24], a new triple-sequence generative adversarial network is presented, consisting of a discriminator, a sentence generator, and an image generator. Image regions for words are generated using the image generator. Based on the generated image regions, the sentence generator is guided by the sentence corpus. We use a large number of unpaired sentences and images to train our model in an unsupervised setting. Experimental results show significant improvements our approach achieves when compared to all baselines.

In [25], a method is presented that leverages zero-shot learning concepts to identify unknown classes and items, combining semantic word embeddings with current state-of-the-art object identification algorithms. Using a pre-trained caption generator, our suggested model, Image Captioning Using New Word Injection, leverages the generator's output to insert objects not in the dataset into the caption. Results outperform the underlying model in both qualitative and quantitative terms.

III. Recognition of Image Caption Generation

A block diagram of the Recognition of Image Caption Generation Using Deep Neural Framework is represented in Figure 1.

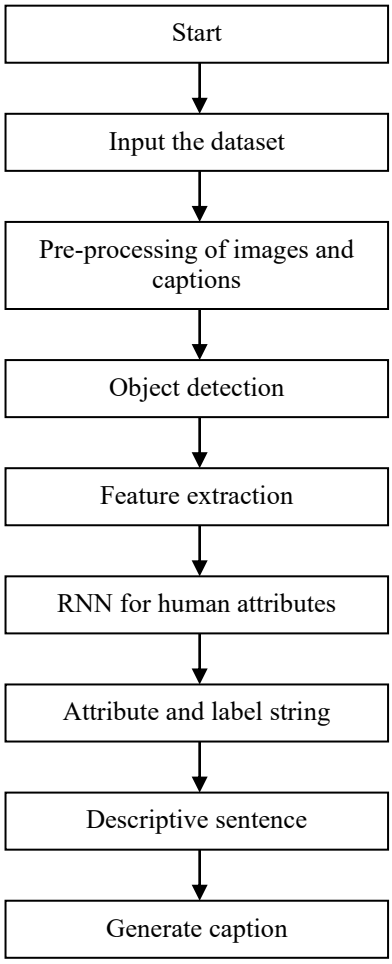

Figure 1: Block diagram of Recognition of Image Caption Generation

The named as"flicker8k" dataset, which we used for our research, work is available online. Flickr8K dataset is castoff meant for image captioning tasks in machine knowledge and CPU vision. It consists of 8,091 images with human-generated captions sourced from the Kaggle website. The dataset contains 40,455 captions, with multiple descriptions per image.

Data Pre-Processing: Initializes an empty dictionary to store the image IDs as keys and agreeing captions as values. Then iterates through each line of caption document, which presents document containing image captions separated by newline characters. Creating tokens which splits each line by commas to separate image ID from captions. Extracts image ID (the first element) and the captions (rest of the elements) from the tokens list. Joins the captions list into a single string by concatenating the captions with a space between them. It checks for the image id in the dictionary if it’s not found then creates new image id then it appends all captions for that image. There is function named clean is a text preprocessing function that takes a mapping dictionary containing image IDs as keys and their corresponding captions as values. The tokeniser class is used to convert to arithmetical sequences suitable for feeding into neural

networks. During this step, the tokenizer assigns a new integer index to each unique word in captions.

Before training the caption generator, we extract features and object properties. The computer system needs to learn the relationships between objects in a given image to understand higher-level meanings. Using NumPy and principal component analysis, the features of an image are extracted. Principal component analysis (PCA) is a statistical technique that reduces the dimensions of data and identifies important features in images.

RNNs are used in this paper for detection, recognition, and caption generation from images. RNNs decode visual features extracted through PCA and produce textual descriptions that corresponding to those features in the context of image captioning. Like feed forward neural networks, the outputs of recurrent neural networks are fed back into input. Networks preserve a hidden state that enables them to change their behaviour during iterations of this feedback loop. Due to their quickly acquire of human grammatical patterns, RNNs are regarded state-of-the-art in machine translation and other text involving jobs. After receiving these features, the RNN creates a caption word by word until either a certain maximum caption length is reached or a new end-of-sequence token is generated.

The model's performance is evaluated using standard evaluation metrics, such as Accuracy, Prediction Time, Loss, and BLEU score.

IV. Result Analysis

The performance of the described Recognition of Image Caption Generation using Deep Neural Framework is presented in this section. It contains 8000 images, of which we used 70% for training and 30% for testing. The image caption model was trained to generate descriptive captions for images. Images were rigorously evaluated using several metrics. Prediction, accuracy, time, loss, and BLEU score were among the evaluation metrics.

A metric called accuracy measures how often a model's predictions are correct. This is determined by dividing the number of correct guesses by the total number of guesses, yielding a percentage ranging from 0% to 100%. The model's predictions are more accurate when they are closer to the actual labels.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad (1)$$

BLEU: Bilingual Evaluation Understudy (BLEU) was initially introduced in the field of machine translation and is used to quantify how closely generated text and reference text match in terms of n-grams, or (n continuous words). The difference between an automatic translation and reference translations of the same source sentence made by humans is measured by BLEU. The precision-based BLEU score ranges from 0 to 1. The closer the value is to 1, the more accurate the closer is.

Prediction Time: It is the Time taken to make predictions or generate captions from input image. If the obtained time is less than it is said to be described model is efficient.

The loss value should be reduced during the training process, the lower the loss value, the good the model.

Table 1 shows the comparative performance of the described Recognition of Image Caption Generation using Deep Neural Framework (RNN) VGG16 (Visual Geometry Group) and ResNet101 models in terms of performance parameters, Accuracy, Prediction Time, Loss and BLEU score.

Table 1: Comparative performance analysis

Parameters	**Image Caption Generation**		
	RNN	**VGG16**	**ResNet101**
Accuracy (%)	98	91	90
Prediction Time (sec)	2	8	9
Loss (%)	6	11	10
BLEU score	0.8	0.5	0.4

Figure 2 shows Accuracy comparative graphical analysis for described Recognition of Image Caption Generation using, Image Caption Generation using Deep Neural Framework (RNN), VGG16 and ResNet101 models. It is found that Accuracy of described model is high compared to other models.

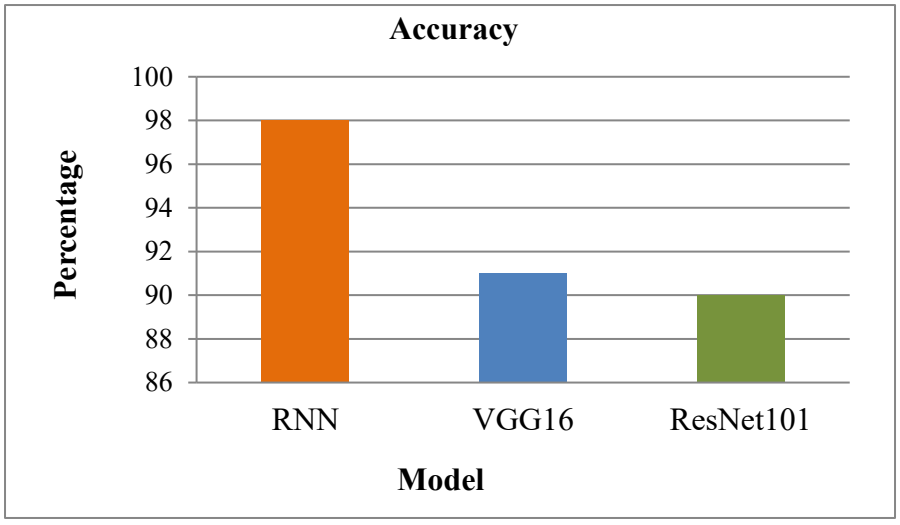

Figure 2: Accuracy parameter comparative performance

The comparative performance of the BLEU score for the described Recognition of Image Caption Generation using Deep Neural Framework (RNN), VGG16, and ResNet101 models is shown in Figure 3, and it is observed that the BLEU score is higher than that of other models.

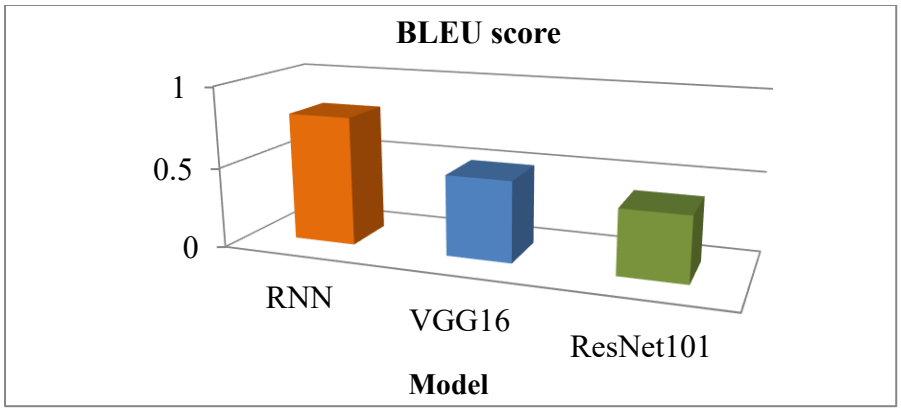

Figure 3: BLEU score comparative analysis

Described Recognition of Image Caption Generation using Image Caption Generation using, Deep Neural Framework (RNN), VGG16 and ResNet101 models comparative Prediction Time parameter is shown in Figure 4. The described model's Prediction Time is very low compared to other models, which declares the efficiency of the model.

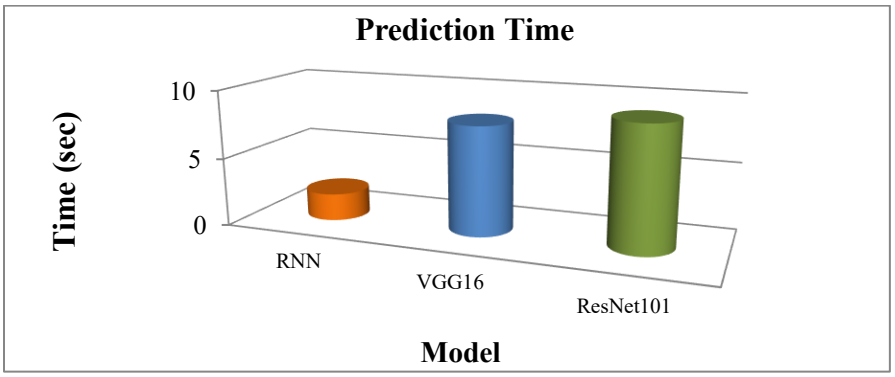

Figure 4: Prediction time analysis

Loss parameter graphical analysis is represented in Figure 5 for the described Recognition of Image Caption Generation using Deep Neural Framework (RNN), VGG16 and ResNet101 models. Loss of the described model is less in percentage, which is really acceptable.

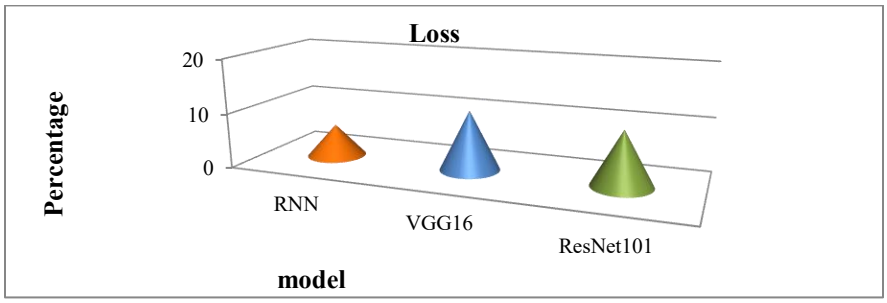

Figure 5: Loss comparative analysis

From the overall analysis, it is observed that the described model, Recognition of Image Caption Generation Using Deep Neural Framework, is efficient in terms of all parameters.

V. Conclusion

Image caption generation using a deep neural framework is described in Recognition. Intelligent human-computer interaction, image search engines, and helping blind and visually impaired people can all be useful applications of automatic image captioning. An 8k Flickr dataset was used to test the proposed approach. RNNs are used in this paper for detection, recognition, and caption generation from images. Principal component analysis is used to extract an image's features. The evaluation metrics included Accuracy, Prediction Time, Loss and BLEU score. From the overall analysis, it is observed that the described model, Recognition of Image Caption Generation Using Deep Neural Framework, is efficient across all parameters.

VI. References

[1] S. Rani S, L. S. Nair and V. M S, "An Enhanced Image Loading Framework for Social Media Applications," *2023 3rd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*, Kalady, Ernakulam, India, 2023, pp. 71-76, doi: 10.1109/ACCESS57397.2023.10200278.

[2] P. Phursutkar and K. Wanjale, "Social re-ranking of image based on visual and semantic information," *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Delhi, India, 2017, pp. 1-6, doi: 10.1109/ICCCNT.2017.8204160.

[3] X. Ye *et al*., "A Joint-Training Two-Stage Method for Remote Sensing Image Captioning," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-16, 2022, Art no. 4709616, doi: 10.1109/TGRS.2022.3224244.

[4] X. Ma, R. Zhao and Z. Shi, "Multiscale Methods for Optical Remote-Sensing Image Captioning," in *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 11, pp. 2001-2005, Nov. 2021, doi: 10.1109/LGRS.2020.3009243

[5] K. Gasmi, H. Aouadi and M. Torjmen, "Link-Driven Study to Enhance Text-Based Image Retrieval: Implicit Links Versus Explicit Links," in *IEEE Access*, vol. 11, pp. 90526-90537, 2023, doi: 10.1109/ACCESS.2023.3307464

[6] Z. Zhang, W. Zhang, W. Diao, M. Yan, X. Gao and X. Sun, "VAA: Visual Aligning Attention Model for Remote Sensing Image Captioning," in *IEEE Access*, vol. 7, pp. 137355-137364, 2019, doi: 10.1109/ACCESS.2019.2942154

[7] T. Wei, W. Yuan, J. Luo, W. Zhang and L. Lu, "VLCA: vision-language aligning model with cross-modal attention for bilingual remote sensing image captioning," in *Journal of Systems Engineering and Electronics*, vol. 34, no. 1, pp. 9-18, February 2023, doi: 10.23919/JSEE.2023.000035

[8] R. Zhao, Z. Shi and Z. Zou, "High-Resolution Remote Sensing Image Captioning Based on Structured Attention," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-14, 2022, Art no. 5603814, doi: 10.1109/TGRS.2021.3070383

[9] A. Ueda, W. Yang and K. Sugiura, "Switching Text-Based Image Encoders for Captioning Images With Text," in *IEEE Access*, vol. 11, pp. 55706-55715, 2023, doi: 10.1109/ACCESS.2023.3282444.

[10] G. Agarwal, K. Jindal, A. Chowdhury, V. K. Singh and A. Pal, "Image and Video Captioning for Apparels Using Deep Learning," in *IEEE Access*, vol. 12, pp. 113138-113150, 2024, doi: 10.1109/ACCESS.2024.3443422.

[11] M. A. Arasi, H. M. Alshahrani, N. Alruwais, A. Motwakel, N. A. Ahmed and A. Mohamed, "Automated Image Captioning Using Sparrow Search Algorithm With Improved Deep Learning Model," in *IEEE Access*, vol. 11, pp. 104633-104642, 2023, doi: 10.1109/ACCESS.2023.3317276

[12] M. Yang *et al*., "Multitask Learning for Cross-Domain Image Captioning," in *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 1047-1061, April 2019, doi: 10.1109/TMM.2018.2869276

[13] G. Sumbul, S. Nayak and B. Demir, "SD-RSIC: Summarization-Driven Deep Remote Sensing Image Captioning," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 8, pp. 6922-6934, Aug. 2021, doi: 10.1109/TGRS.2020.3031111.

[14] G. Bharathi Mohan, R. Harigaran, P. Sri Varshan, R. Srimani, R. Prasanna Kumar and R. Elakkiya, "Image Caption Generation using Contrastive Language Image Pretraining," *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India, 2024, pp. 1-5, doi: 10.1109/ICCCNT61001.2024.10725907.

[15] Y. Yang, H. Wei, H. Zhu, D. Yu, H. Xiong and J. Yang, "Exploiting Cross-Modal Prediction and Relation Consistency for Semisupervised Image Captioning," in *IEEE Transactions on Cybernetics*, vol. 54, no. 2, pp. 890-902, Feb. 2024, doi: 10.1109/TCYB.2022.3156367.

[16] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga and M. Bennamoun, "Text to Image Synthesis for Improved Image Captioning," in *IEEE Access*, vol. 9, pp. 64918-64928, 2021, doi: 10.1109/ACCESS.2021.3075579.

[17] Q. Wang, W. Huang, X. Zhang and X. Li, "Word–Sentence Framework for Remote Sensing Image Captioning," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 12, pp. 10532-10543, Dec. 2021, doi: 10.1109/TGRS.2020.3044054.

[18] C. Wu, S. Yuan, H. Cao, Y. Wei and L. Wang, "Hierarchical Attention-Based Fusion for Image Caption With Multi-Grained Rewards," in *IEEE Access*, vol. 8, pp. 57943-57951, 2020, doi: 10.1109/ACCESS.2020.2981513.

[19] M. Al-Qatf, X. Wang, A. Hawbani, A. Abdussalam and S. H. Alsamhi, "Image Captioning With Novel Topics Guidance and Retrieval-Based Topics Re-Weighting," in *IEEE Transactions on Multimedia*, vol. 25, pp. 5984-5999, 2023, doi: 10.1109/TMM.2022.3202690

[20] S. Cao, G. An, Z. Zheng and Z. Wang, "Vision-Enhanced and Consensus-Aware Transformer for Image Captioning," in *IEEE Transactions on Circuits and Systems for*

Video Technology, vol. 32, no. 10, pp. 7005-7018, Oct. 2022, doi: 10.1109/TCSVT.2022.3178844

[21] L. Wu, M. Xu, L. Sang, T. Yao and T. Mei, "Noise Augmented Double-Stream Graph Convolutional Networks for Image Captioning," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 8, pp. 3118-3127, Aug. 2021, doi: 10.1109/TCSVT.2020.3036860

[22] S. -H. Han and H. -J. Choi, "Domain-Specific Image Caption Generator with Semantic Ontology," *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, Busan, Korea (South), 2020, pp. 526-530, doi: 10.1109/BigComp48618.2020.00-12.

[23] D. -H. Hoang, A. -K. Tran, D. N. Minh Dang, P. -N. Tran, H. Dang-Ngoc and C. T. Nguyen, "RBBA: ResNet - BERT - Bahdanau Attention for Image Caption Generator," *2023 14th International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju Island, Korea, Republic of, 2023, pp. 430-435, doi: 10.1109/ICTC58733.2023.10392496.

[24] Y. Zhou, W. Tao and W. Zhang, "Triple Sequence Generative Adversarial Nets for Unsupervised Image Captioning," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, 2021, pp. 7598-7602, doi: 10.1109/ICASSP39728.2021.9414335.

[25] M. M. A. Baig, M. I. Shah, M. A. Wajahat, N. Zafar and O. Arif, "Image Caption Generator with Novel Object Injection," *2018 Digital Image Computing: Techniques and Applications (DICTA)*, Canberra, ACT, Australia, 2018, pp. 1-8, doi: 10.1109/DICTA.2018.8615810.